Quantifying kinetics from time series of single-molecule Förster resonance energy transfer efficiency histograms

Benke, Stephan ; Nettels, Daniel ; Hofmann, Hagen ; Schuler, Benjamin

Abstract: Single-molecule fluorescence spectroscopy is a powerful approach for probing biomolecular structure and dynamics, including protein folding. For the investigation of nonequilibrium kinetics, Förster resonance energy transfer combined with confocal multiparameter detection has proven particularly versatile, owing to the large number of observables and the broad range of accessible timescales, especially in combination with rapid microfluidic mixing. However, a comprehensive kinetic analysis of the resulting time series of transfer efficiency histograms and complementary observables can be challenging owing to the complexity of the data. Here we present and compare three different methods for the analysis of such kinetic data: singular value decomposition, multivariate curve resolution with alternating least square fitting, and model-based peak fitting, where an explicit model of both the transfer efficiency histogram of each species and the kinetic mechanism of the process is employed. While each of these methods has its merits for specific applications, we conclude that model-based peak fitting is most suitable for a quantitative analysis and comparison of kinetic mechanisms.

DOI: https://doi.org/10.1088/1361-6528/aa5abd

Quantifying Kinetics from Time Series of Single-Molecule FRET Efficiency Histograms

Stephan Benke[1], Daniel Nettels[1], Hagen Hofmann[1,2] and Benjamin Schuler[1,3,4]

1: Department of Biochemistry and 3: Department of Physics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
2: Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel
4: Corresponding author: schuler@bioc.uzh.ch

Abstract

Single-molecule fluorescence spectroscopy is a powerful approach for probing biomolecular structure and dynamics, including protein folding. For the investigation of nonequilibrium kinetics, Förster resonance energy transfer (FRET) in combination with confocal multiparameter detection has proven particularly versatile, owing to the large number of observables and the broad range of accessible timescales, especially in combination with rapid microfluidic mixing. However, a comprehensive kinetic analysis of the resulting time series of transfer efficiency histograms and complementary observables can be challenging owing to the complexity of the data. Here we present and compare three different methods for the analysis of such kinetic data: singular value decomposition, multivariate curve resolution with alternating least square fitting, and model-based peak fitting, where an explicit model of both the transfer efficiency histogram of each species and the kinetic mechanism of the process is employed. While each of these methods has its merits for specific applications, we conclude that model-based peak fitting is most suitable for a quantitative analysis and comparison of kinetic mechanisms.

1. Introduction

Single-molecule methods have developed into essential tools for investigating biomolecular processes. A particularly versatile approach is the combination of single-molecule fluorescence detection with Förster resonance energy transfer [8] (FRET), which enables distances and distance changes on the nanometre length scale to be monitored with high sensitivity [41, 42]. Single-molecule FRET provides a powerful way of resolving heterogeneity, *e.g.*, the conformational heterogeneity arising in protein folding and misfolding reactions, and the occurrence of rare events that would elude detection by ensemble-averaging techniques [39].

However, the method not only provides access to equilibrium distributions but also to the kinetics of biomolecular reactions. In many cases, dynamics and kinetics can be monitored at equilibrium by virtue of the spontaneous fluctuations occurring at the single-molecule level [41]. For example, the time series of signals from FRET-labelled molecules immobilized on a surface are commonly evaluated using dwell time analysis [38], hidden Markov models [31], or maximum likelihood techniques [11]. These approaches often give detailed insights into the timescales of the process, the existence of different conformational states, and the connectivity between them, thus providing a powerful way of kinetic model building. Similarly, confocal fluorescence detection of freely diffusing molecules can provide information on the equilibrium dynamics of biomolecules on timescales from nanoseconds to milliseconds with approaches such as correlation spectroscopy [37, 39], detailed modelling of the photon statistics [1, 12, 33], or recurrence analysis [18]. However, all these methods require the states or conformations of interest to be sufficiently populated at equilibrium or under steady-state conditions; at the same time, they require the kinetics of interconversion to be accessible on the timescale of the

fluorescence recording from an individual molecule. These combined requirements are often not met by biomolecular systems, whose dynamics can cover a vast range of timescales, from nanoseconds to days or weeks [39].

An alternative approach for the investigation of biomolecular dynamics is to use single-molecule detection jointly with perturbation techniques, such as rapid mixing or laser-induced temperature jumps. For example, experiments that combine microfluidic and manual mixing techniques with single-molecule fluorescence spectroscopy give access to non-equilibrium kinetics of the structural changes in proteins (Fig. 1). This approach is similar to established ensemble-based kinetic techniques, but with the advantage of obtaining distributions instead of average signals, and (provided suitable detection systems are available) for multiple parameters [42], such as FRET efficiencies, fluorescence lifetimes, anisotropies, and correlation functions. By monitoring single molecules free in solution, the time evolution of samples can be monitored for hours or days [5], and microfluidic mixing [27, 30] reduces dead times to milliseconds [30, 35, 45] and below [10]. The transient population of metastable states enables the detailed analysis of their dynamics down to nanoseconds even if they are not detectable at equilibrium [4, 43]. In this way, the kinetic mechanisms of complex reactions involving up to half a dozen species have been elucidated [2, 3]. However, taking full advantage of the rich information available from these experiments in a self-consistent manner calls for advanced analysis techniques.

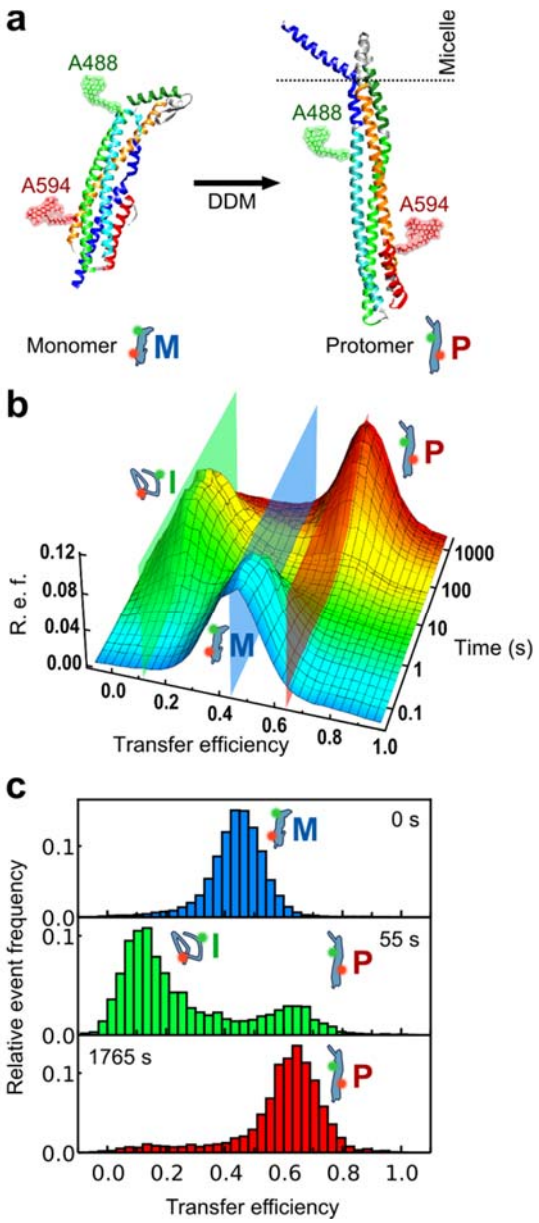

Figure 1. The protomer formation process of the pore-forming toxin ClyA followed by single-molecule spectroscopy as an example for a transfer efficiency histogram time series. (**a**) Upon binding to membranes (or in this case n-dodecyl-β-D-maltopyranoside (DDM) micelles), ClyA undergoes a large conformational transition from the soluble monomer state to the membrane-bound protomer state. This process can be monitored with single-molecule FRET. (**a**) Structures of the monomer and the pore conformation of ClyA (PDB code 1QOY [44], PDB code 2WCD [32]). The protomer conformation is represented by one pore subunit. A488 and A594 and the atomic dye structures indicate the positions labelled with Alexa Fluor 488 and 594, respectively. (Structure representations created with Chimera [34] and Avogadro [14].) (**b**) Time series of area-normalized transfer efficiency histograms measured for the monomer-to-protomer transition in 0.1% (w/v) DDM [2]. The cartoons illustrate the monomer (M), intermediate (I) and protomer (P), and the transparent planes indicate the average transfer efficiencies of the peaks corresponding to the individual subpopulations. R. e. f.: Relative event frequency. (**c**) Individual histograms from the time series in (b) at times where one individual state dominates. For each histogram, the time is given in seconds, and the cartoons indicate the conformational states. Figure adapted from Ref. [2].

Here we present and review several approaches that allow a detailed kinetic analysis of FRET efficiency histogram time series covering times from milliseconds to hours: multivariate curve resolution with alternating least square fitting (MCR-ALS) [7, 21, 22], singular value decomposition (SVD) [16, 47], and model-based peak fitting [2, 7]. SVD can provide a virtually model-free assessment of the number of states or kinetic components involved in a process of interest. MCR-ALS is a technique that enables

us to estimate the signal characteristics and time courses of individual species involved in a process. The underlying kinetic mechanism can often not be identified directly from an SVD or MCR-ALS analysis. For this purpose, a description of the data with models for both the signal characteristics of the species and the kinetics of interconversion is usually necessary. We will illustrate the approach of explicitly modelling the kinetics of single-molecule FRET efficiency histograms and show how it can be informed and supplemented by model-free methods such as MCR-ALS and SVD. Further, we will show how to incorporate additional observables available in multiparameter single-molecule fluorescence detection.

2. Matrix representation of transfer efficiency histogram time series

The experimental data from kinetic single-molecule measurements in free diffusion can be represented as a series of N transfer efficiency histograms. The histograms are recorded at different times t_n ($n = 1, \dots N$) after the start of a reaction, *e.g.*, when triggered by rapid mixing. For a kinetic analysis, we want to obtain the time course of the population of every species observed in the histograms. Each histogram of the time series can be written as a vector:

$$\boldsymbol{h}_n = \begin{pmatrix} h_{1n} \\ \vdots \\ h_{Mn} \end{pmatrix}$$

Here, h_{mn} with $m = 1, \dots M$ is the number of burst events, each originating from an individual molecule diffusing through the confocal observation volume, with transfer efficiency values, E, that satisfy $E_m - \Delta/2 < E \leq E_m + \Delta/2$, where Δ is the bin width in the transfer efficiency histogram, and E_{m} is the midpoint transfer efficiency of the mth bin. We assume that each histogram can be represented as a linear combination of 'species histograms', $\boldsymbol{f}_l$, *i.e.*, the transfer efficiency histograms corresponding to the individual species involved in the process:

$$\boldsymbol{h}_n = \sum_{l=1}^{L} c_{nl}\, \boldsymbol{f}_l,$$

where

$$\boldsymbol{f}_l = \begin{pmatrix} f_{1l} \\ \vdots \\ f_{Ml} \end{pmatrix} \text{ with } \sum_{m=1}^{M} f_{ml} = 1.$$

The species histograms would result from measurements of samples containing only one of the species $l = 1, \dots L$ (which in practice, however, is often difficult to obtain, as discussed below).

$$c_{nl} = c_l(t_n)$$

is the concentration of species l at time t_{n} (can also be expressed relative to the respective starting concentration c_{l0} if desired).

For the subsequent analysis, we combine all histograms $\boldsymbol{h}_n$ of the time series in one M x N matrix $\boldsymbol{H} = (\boldsymbol{h}_1, \dots \boldsymbol{h}_N)$ and, likewise, all species histograms f_l in the M x L matrix $\boldsymbol{F} = (\boldsymbol{f}_1, \dots \boldsymbol{f}_L)$. The relative concentrations, c_{nl}, define an additional N x L matrix:

$$\boldsymbol{C} = \begin{pmatrix} c_{11} & \cdots & c_{1L} \\ \vdots & \ddots & \vdots \\ c_{N1} & \cdots & c_{NL} \end{pmatrix}.$$

With these definitions, we can write:

$$\boldsymbol{H} = \boldsymbol{F}\boldsymbol{C}^T.$$

The aim of our analysis is to find the decomposition of $\boldsymbol{H}$ into $\boldsymbol{F}$ and $\boldsymbol{C}$. If $\boldsymbol{F}$ is already known, *e.g.*, from measurements of samples of the individual species, we can obtain $\boldsymbol{C}$ from linear regression by minimizing

$$\chi^2 = \|\boldsymbol{H} - \boldsymbol{F}\boldsymbol{C}^T\|_F^2,$$

where $\|A\|_F{}^2 = \sum_{i,j} \left|a_{ij}\right|^2$ is the Frobenius norm. In practice, however, it is often not possible to isolate all species to obtain the individual species histograms. Additionally, the peaks of the subspecies frequently overlap, which can lead to ambiguities in $\boldsymbol{C}$. The methods reviewed and presented here are designed to extract as much information on $\boldsymbol{F}$ and $\boldsymbol{C}$ from $\boldsymbol{H}$ as possible under these suboptimal conditions.

3. Multivariate curve resolution – alternating least square fitting (MCR-ALS)

With the MCR-ALS method [7, 21, 22], $\boldsymbol{F}$ and $\boldsymbol{C}$ are obtained by an iterative process in which χ^2 is alternatingly minimized with respect to $\boldsymbol{F}$ or $\boldsymbol{C}$ while the respective other matrix is kept fixed. This alternating minimization is repeated until the fit reaches convergence. To start the process, an initial guess is required, either for $\boldsymbol{F}$ or $\boldsymbol{C}$. Usually, several constraints can be imposed on $\boldsymbol{F}$ and $\boldsymbol{C}$. In the case of FRET efficiency histogram time series, both matrices are non-negative, *i.e.*, $f_{ml} \geq 0$ and $c_{nl} \geq 0$ (i.e., requiring non-negative matrix factorization [29] of $\boldsymbol{H}$). In addition, the total concentration of species is constant over time in many cases:

$$\sum_{l=1}^{L} c_{nl} = c_0 \text{ for all } n = 1, \dots N.$$

By definition, the species histogram vectors are normalized, *i.e.*,

$$\sum_{m=1}^{M} f_{ml} = 1 \text{ for all } l = 1 \dots L.$$

Furthermore, species histograms are typically unimodal:

$$f_{m-1\,l} \leq f_{ml} \text{ for } m \leq m_0$$
$$f_{ml} \geq f_{m+1\,l} \text{ for } m \geq m_0,$$

where m_0 is the histogram bin at the peak maximum.

MCR-ALS is illustrated in Fig. 2 for the large conformational change that occurs during the formation of the protomer of cytolysin A (ClyA) upon rapid mixing of the monomer with dodecyl maltoside (DDM) [2]. For the analysis, we assumed the presence of three different species (monomer (M), intermediate (I), and protomer (P)) in the reaction and used the relative peak heights over time as an initial guess for $\boldsymbol{C}$. Displayed are three results using different constraints. The first analysis was done using the GUI MCR-ALS implementation in Matlab by Jaumot *et al*. [23] and employing only the constraints described above. In many situations, such a virtually model-free analysis (only the number of species is imposed) can already provide a reasonable estimate of the species histograms, and, provided that the individual species are resolved sufficiently, even of their time evolution.

For the second and third analysis, we constrained the concentrations c_{nl} with specific kinetic models, in this case with an off-pathway model and an on-pathway model, as depicted in Fig. 2b. The use of models to describe one of the matrices ($\boldsymbol{F}$ or $\boldsymbol{C}$) while the other is determined by MCR-ALS has been termed 'combined hard and soft modelling' (HSM) [7] and is available in the latest GUI MCR-ALS Matlab implementation by Tauler and co-workers [21]. In the following, we use the off-pathway model to illustrate how one can constrain $\boldsymbol{C}$ to solutions of a kinetic model. The corresponding rate equations for the monomer (c_M), intermediate (c_I), and protomer (c_P) concentrations are:

$$\frac{d\,c_I(t)}{dt} = k_1\,c_M(t) - k_2\,c_I(t),$$

$$\frac{d\,c_M(t)}{dt} = k_2\,c_I(t) + k_4\,c_P(t) - (k_1 + k_3)\,c_M(t), \text{ and}$$

$$\frac{d\,c_P(t)}{dt} = k_3\,c_M(t) - k_4\,c_P(t),$$

with starting conditions $c_M(0) = 1$, $c_I(0) = 0$, and $c_P(0) = 0$. The system of equations is solved numerically for a given set of rate coefficients k_1 to k_4, and the concentrations of the three species are calculated for all times t_n to obtain the matrix $\boldsymbol{C}$. The rate coefficients are then optimized iteratively using common algorithms for minimizing χ^2 for the entire series of transfer efficiency histograms.

The use of a kinetic model to describe $\boldsymbol{C}$ has two advantages. First, the kinetic model helps to resolve ambiguities in the concentrations of the species in the case of strongly overlapping histogram peaks, and second, it drastically reduces the number of parameters needed to describe $\boldsymbol{C}$ because only the rate coefficients and the starting conditions enter into the model. A comparison between the analysis with and without constraints by the kinetic models is shown in Fig. 2c. Although the resulting pairs of $\boldsymbol{F}$ and $\boldsymbol{C}$ matrices differ substantially for the two kinetic models, the reconstructed histogram time series, $\boldsymbol{FC}^T$, are nearly identical and result in similar χ^2 values (Fig. 2d). The on-pathway model yields the lowest $\chi 2$, but the species histogram of the intermediate shows extensive peak tailing towards high transfer efficiency (Fig. 2c), which is not expected for a pure species corresponding to a narrow range of distances or rapidly interconverting conformations [12] (Fig. 1c). The on-pathway model thus provides a (slightly) better fit than the off-pathway model but yields more complex species histograms in $\boldsymbol{F}$ that correspond to mixtures of the actual species involved. This example illustrates some of the strengths and weaknesses of the approach. Using the MCR-ALS method to fit kinetic models to these data can provide a reasonable decomposition of the transfer efficiency time series, especially if the peaks from different subpopulations are well separated. However, since the shapes of the species histograms and the rate coefficients are not completely independent parameters in the analysis, mutual compensation can occur that leads to ambiguity in the results. In some cases, this may result in obviously unphysical species histograms, but in more subtle scenarios such as the one illustrated here, it poses the problem of choosing one set of species histograms over another and thus requires additional information on the pure species.

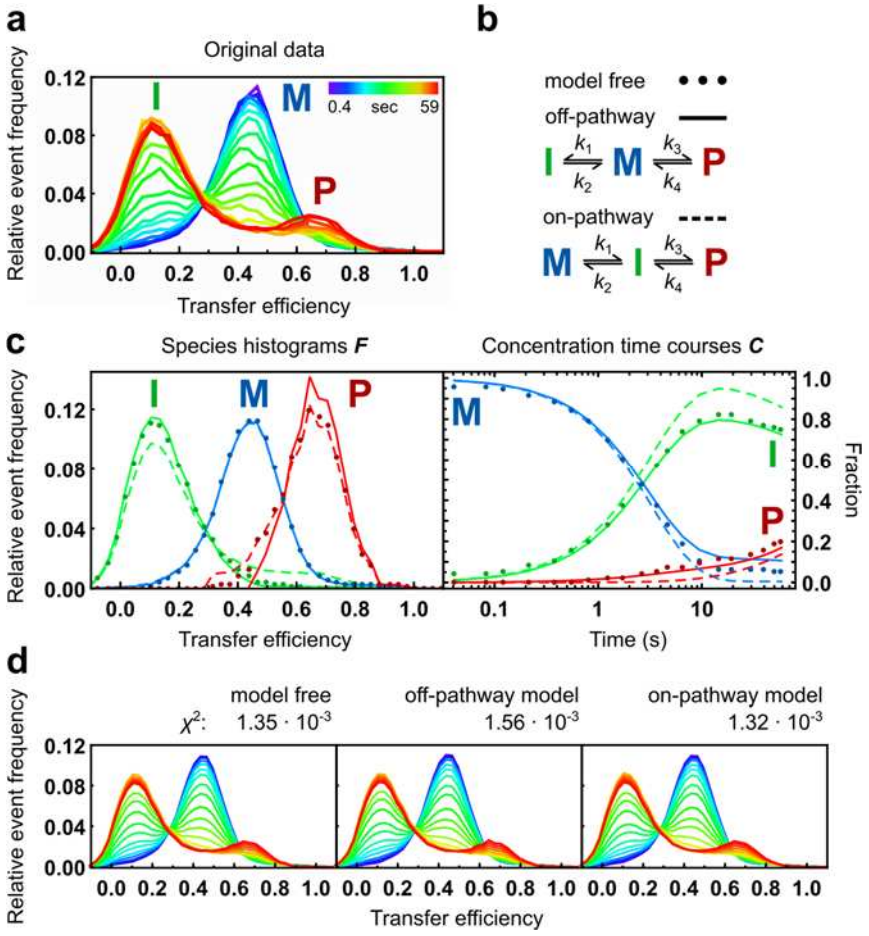

Figure 2. Example of MCR-ALS applied to a dataset. (**a**) Measured transfer efficiency histograms (normalized, $\sum_m h_{mn} = 1$) for the first minute of ClyA protomer formation in 0.1% DDM [2]. Each line represents one transfer efficiency histogram, with the associated time after start of the reaction color-coded (see legend). (**b**) The data were analysed by MCR-ALS in three ways: no kinetic model describing the matrix of species concentrations, $\boldsymbol{C}$, (model-free), and $\boldsymbol{C}$ described by an off- or an on-pathway kinetic model. (**c**) Species histograms (left) and their concentration time courses (right) resulting from the three fits with the models in (b). The coloured points represent the results from the free fit, solid lines from the off-pathway model, and dashed lines from the on-pathway model. (**d**) Reconstruction of the data ($\boldsymbol{FC}^T$) from the three different types of analysis.

4. Model-based peak fitting

Constraining the shapes of the species histograms by realistic model functions can help to avoid the ambiguities in the results described in the previous section. Most rigorously, model species histograms can be derived from a detailed theoretical description of the underlying photon statistics, especially shot noise, of fluorescent species diffusing through the confocal volume. Gopich and Szabo used such an approach to derive analytical functions to model FRET efficiency histograms [12, 13]. Photon distribution analysis (PDA) is a related approach where the experimentally observed burst size distribution is included to model transfer efficiency distributions [1, 24, 33]. However, in many cases, FRET efficiency histograms can be well approximated with sums of simple peak functions to obtain reliable results on the relative populations of the constituting species without the need for detailed information on the underlying photon statistics. In addition, such simple functions increase the computational speed substantially, which is particularly advantageous for the analysis of large data sets. Even though a full incorporation of photon statistics is possible, we thus focus on an approach using peak functions based on simple analytical expressions.

Commonly used are Gaussian peak functions for symmetric peaks and four-parameter log-normal peak functions for skewed peaks since they suitably reproduce the peak shapes dictated by the underlying photon statistics [17, 28, 40, 46]. Asymmetric peaks are observed for species with high or low mean FRET efficiencies [6, 12]. With the Gaussian and log-normal peak functions, the elements f_{ml} of the species histograms, $\boldsymbol{f}_l$, are then given by

$$f_{ml} = A_l \exp\left[-\frac{(E_m - e_l)^2}{2\,w_l^2}\right]$$

for symmetric peaks and by

$$f_{ml} = A_l \exp\left[-\frac{\ln 2}{\ln(a_l)^2}\left\{\ln\left(1 + \frac{{a_l}^2 - 1}{a_l w_l}(E_m - e_l)\right)\right\}^2\right]$$

for asymmetric peaks. Here, E_m is the midpoint transfer efficiency of the mth bin, e_l is the peak position, and w_l and a_l are the width and the asymmetry of the peak for species l, respectively. The normalization factor, A_l, is chosen in order that $\sum_{m=1}^{M} f_{ml} = 1$. Examples of the peak functions are shown in Fig. 3.

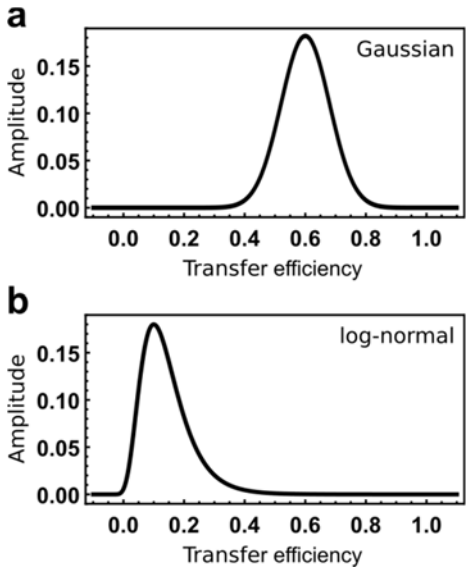

Figure 3. Empirical peak functions used to describe transfer efficiency histograms. (**a**) Example of a Gaussian peak function for symmetric peaks with $e_l = 0.6$ and $w_l = 0.08$. (**b**) Example of a four-parameter log-normal peak function for asymmetric peaks with $e_l = 0.1$, $w_l = 0.15$, and $a_l = 1.6$.

In favourable cases, the parameters defining the shapes and positions of the peak functions (w_l, a_l, e_l) can be determined independently by fitting the histograms of the pure species populated exclusively or well separated from other populations under suitable conditions. In the analysis of the complete histogram time series, these parameters can then be fixed, leaving only the relative species concentrations as adjustable fit parameters. Alternatively, especially for species for which no reliable individual species histograms can be obtained, the peak parameters can often be determined consistently by fitting the entire histogram time series with peak functions, but with the unknown parameters e_l, w_l, and a_l as shared global fit parameters, i.e., by requiring that the each species is described by a single set of peak parameters across the entire data set. Those peak parameters that could be determined independently can be fixed to the known values, thus leaving only the amplitudes A_l unconstrained and independent for all histograms. This approach is robust if each species is clearly populated in the histograms at some time during the reaction and if the species histograms are not excessively broad, overlapping, or changing in mean transfer efficiency or width as a function of time. Note that although

the peak shapes obtained in this way are independent of any kinetic model, the resulting peak areas can already provide a good idea of the changes in species concentrations as a function of time. In a next step, the changes in concentrations are then modelled explicitly in terms of a detailed kinetic scheme

Figure 4 illustrates the complete fit procedure for the same data set as shown in Fig. 2. In a first step, the measured histogram time series (Fig. 4a) is fitted with three peak functions for the three populations (I, M, and P) that are readily identifiable: I with a log-normal peak function, and M and P with Gaussian peak functions (Fig. 4b). Since the peaks are well separated and all species are sufficiently populated at some time during the reaction, we obtain the relevant peak parameters reliably from a global fit to the entire set of histograms, thus yielding, after normalization, the species histograms in $\boldsymbol{F}$. In cases where a species is barely populated or where its peak strongly overlaps with that of another species, its peak parameters have to be obtained by separate measurements whenever possible. Using these descriptions of the species histograms, a kinetic model (in this example the off-pathway model, Fig. 2b) can then be used to fit the data in the same way as described for MCR-ALS. Since the peak function parameters were determined independent of the kinetics, the only remaining adjustable parameters are the rate coefficients of the kinetic model. The resulting fit yields the species concentration time courses contained in matrix $\boldsymbol{C}$ (Fig. 4c). $\boldsymbol{F}$ and $\boldsymbol{C}$ then allow the data to be reconstructed (Fig. 4d). The corresponding residuals, $\boldsymbol{H} - \boldsymbol{F}\boldsymbol{C}^T$, are shown in Fig. 4e.

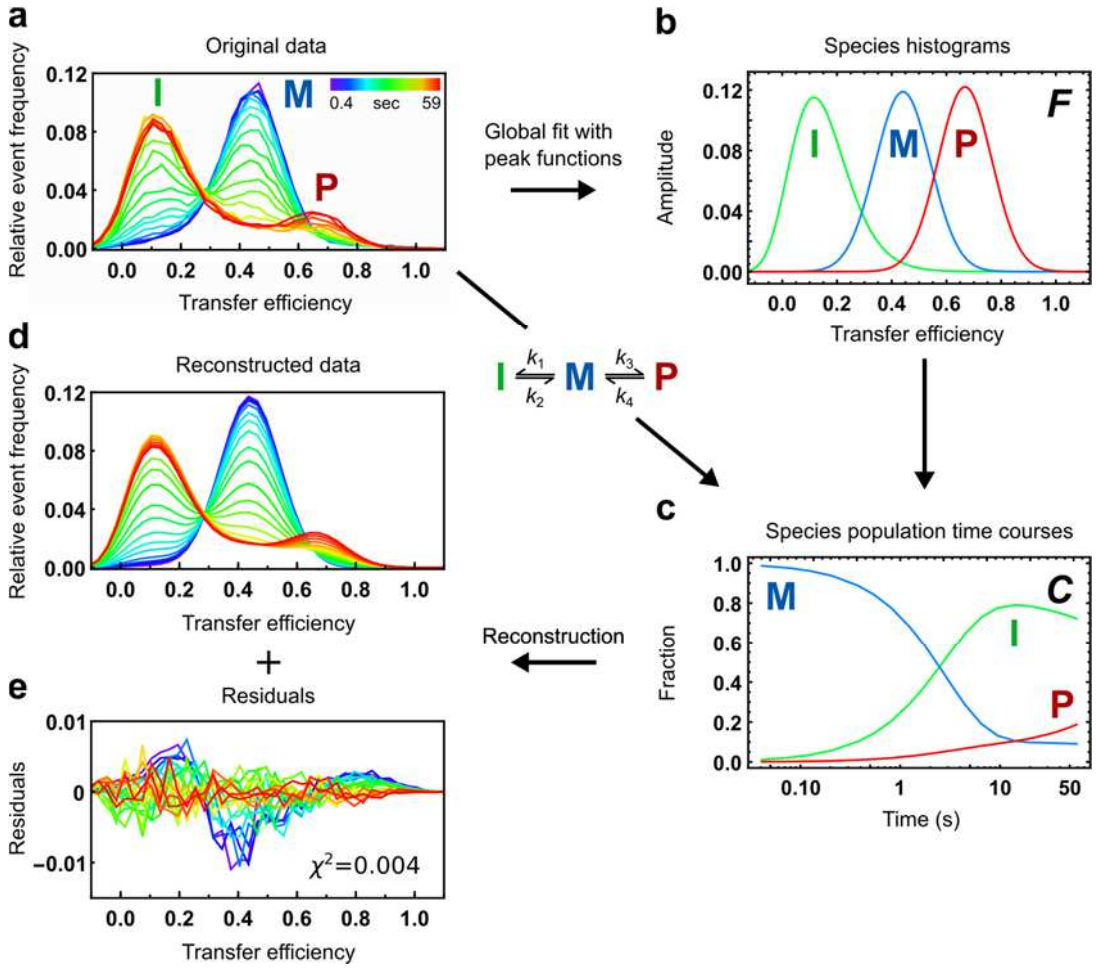

Figure 4. Fitting a transfer efficiency histogram time series with a combination of peak functions and a kinetic model. (a) Measured transfer efficiency histogram time series (see Fig. 1a for details.). Three different peaks, one at low (intermediate I), one at intermediate (monomer M), and one at high (protomer P) transfer efficiency can be distinguished. (b) Species histograms (matrix F) resulting from a fit of the time series in (a) with one log-normal (for I) and two Gaussian peak functions (for M and P), with peak position el, widths wl, and asymmetries al as global fit parameters. (c) Population time courses of M, I, and P (matrix C) resulting from a fit of the histogram time series in (a) with the species histograms in (b) and the off-pathway

kinetic model. (d) Histogram time series reconstructed from the species histograms (b) and species population time courses (c). (e) Residuals of the fit. χ^2 is the total sum over all squared residuals.

With the species histograms, $\boldsymbol{F}$, determined independently of $\boldsymbol{C}$, the procedure described here allows a direct comparison of different kinetic models, in our case the off- and on-pathway models (Fig. 5). The quality of the fits of the two models can be visualized and compared on different levels. For the individual histograms, the residuals of the fits can be plotted (Fig. 5a). To assess the fit quality over the entire time course of the reaction, the residuals of each histogram can be squared, and the sum of squared residuals ($SSR = \sum_m (h_{mn} - \sum_l c_{nl} f_{ml})^2$) of each histogram can be plotted against t_n (Fig. 5b). Particularly instructive is the ratio of the SSR for the two models as function of time (Fig. 5c). In the present example, the SSR ratios clearly show that the off-pathway model provides a markedly better fit to the data between 1 – 100 s, the very interval in which the predicted concentration time courses differ the most (Fig. 5b). The differences between the two models may appear small at the level of the individual histograms, but they are clearly identified by the SSR ratio.

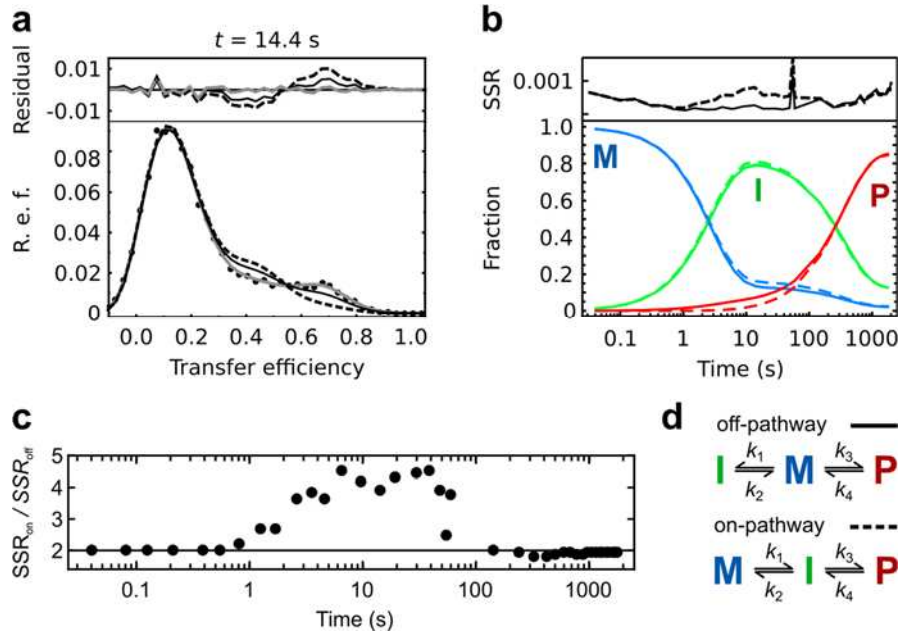

Figure 5. Comparison of the fit quality for different kinetic models. (**a**) Visualization of the fit quality of different models on the level of a single histogram. In the lower panel, the black dots represent the measured histogram (in this case at t = 14.4 s), and the solid and dashed lines show the fit to the data using either no kinetic model (solid grey line), the off-pathway model (solid black line), and the on-pathway model (dashed black line). The upper panel shows the residuals. (**b**) Model fit quality over the course of the complete dataset. The concentration time courses for the monomer (M, blue), intermediate (I, green), and protomer (P, red) (***C***) resulting from the fits with the two different models (d) are shown in the lower panel. The upper panel shows the corresponding sums of squared residuals (SSRs, see main text). (**c**) SSRs for the on-pathway fit divided by those of the off-pathway fit plotted over the time course of the reaction. The increase in the SSR ratios between 1 and 100 s illustrates the better fit of the off-pathway model. (**d**) The off- and on-pathway kinetic models used for the fits. Data taken from [2].

The robustness of the procedure and the results can be tested in several ways, e.g., by bootstrapping or by a systematic variation of the peak function parameters used in the fit. For bootstrapping, a number of synthetic data sets are produced by randomly sampling with replacement from the measured sets of photon bursts. Subsequently, these data sets are analysed in the same way as the original data [36]. Bootstrapping allows realistic estimates to be obtained for the statistical uncertainties of the fit results related to the analysis procedure, excluding systematic errors. To simulate reduced data quality, the new data sets can be constructed from a reduced number of photon bursts, or noise can be added to the transfer efficiency of each burst. The influence of the uncertainty in the shapes and positions of the peak functions on the rate coefficients can be tested by systematic variation of the peak function parameters.

Both bootstrapping and peak parameter variation can identify the relative uncertainty of the fit parameters and illustrate how robustly kinetic models can be distinguished. In our example (Fig. 5), the kinetic off-pathway model consistently resulted in smaller χ^2 values than the on-pathway model.

We note, however, that systematic errors can result from the inadequacy of simple peak functions for describing peak tailing or other aberrant histogram shapes resulting from sample heterogeneity or photochemical effects such as photobleaching. Optimizing sample quality and experimental conditions to minimize such effects is thus essential for quantitative analysis. In the case of static heterogeneity on the timescale of observation (see *Multiparameter singular value decomposition* for an example), simple peak functions may be inadequate, and a more detailed analysis of the underlying conformational distributions and dynamics may be required [18, 20]. Finally, it is worth pointing out again that the methods used here can in principle be combined with a detailed description of the transfer efficiency histograms in terms of the underlying photon statistics [1, 12, 33], but for more complex kinetic mechanisms, the computational cost of a global analysis may still be a challenge for routine use.

5. Singular value decomposition (SVD)

A method that is frequently used for the analysis of multivariate experimental data and that is largely model-free is singular value decomposition (SVD) [36]. It has recently also started to be employed for the analysis of single-molecule experiments [9, 19, 20]. The procedures for decomposing $\boldsymbol{H}$ into $\boldsymbol{F}$ and $\boldsymbol{C}$ presented above require *a priori* knowledge of the number of species present in the reaction and, additionally, information on either the shapes of their individual *E* histograms or on the time courses of their relative concentrations. In contrast, SVD allows a model-free analysis of the data that provides information about the number of distinguishable species involved in the reaction and, in addition, where and on which time scales changes occur in the signals [16].

However, SVD does not yield a decomposition of $\boldsymbol{H}$ into $\boldsymbol{F}$ and $\boldsymbol{C}$. Instead, it expresses the *M x N* matrix $\boldsymbol{H}$ as a product of three matrices:

$$\boldsymbol{H} = \boldsymbol{USV}^T.$$

$\boldsymbol{U}$ is an orthonormal *M x M* matrix whose columns, the left singular vectors of $\boldsymbol{H}$, contain the information about the shapes of the transfer efficiency histograms. $\boldsymbol{S}$ is an *M x N* diagonal matrix that contains the singular values that determine to which extent each pair of columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, *i.e.*, the components of the SVD, contributes to $\boldsymbol{H}$. The singular values are sorted by magnitude, *i.e.*, the largest singular value is the first diagonal element of $\boldsymbol{S}$. The columns of the orthonormal *N x N* matrix $\boldsymbol{V}$, the right singular vectors or amplitude vectors, contain the kinetics. When weighted by their singular value, they provide the relative contribution of the corresponding columns of $\boldsymbol{U}$ at the different times t_n.

5.1. SVD for FRET efficiency histogram time series

We illustrate SVD for the same data for ClyA [2] used above (Fig. 6). The measured transfer efficiency histograms, which clearly show the presence of three distinguishable species (Fig. 6a), are decomposed into the shape components (columns of $\boldsymbol{U}$, Fig. 6e), the weights (diagonal elements of $\boldsymbol{S}$, Fig. 6c), and the kinetics (columns of $\boldsymbol{V}$, Fig. 6f). To determine the minimum number of species in the histogram time series, two measures are indicative: the singular values in $\boldsymbol{S}$, and the autocorrelations of the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ (see below). The relative magnitudes of the singular values provide an estimate of the relevant sets of vectors in $\boldsymbol{U}$ and $\boldsymbol{V}$ and which of the vectors are likely to contain only noise. In our example, the

first three singular values are substantially greater than the remaining ones, indicating three significant species (Fig. 6c). However, the difference between the third and fourth singular values does not suffice to unequivocally decide on the number of species. As an additional measure, we can examine the variations in the left and right singular vectors by means of the autocorrelations, *G*, of the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ [16]:

$$G(U_j) = \sum_{i=1}^{M-1} U_{i,j} U_{i+1,j}$$

$$G(V_j) = \sum_{i=1}^{N-1} V_{i,j} V_{i+1,j}$$

Possible values of *G* range from 1 (complete correlation) to -1 (no correlation), and values above 0.8 are considered to have a high signal-to-noise ratio for large vectors [16]. For transfer efficiency histograms and the kinetics on timescales considered here, we found the autocorrelations to be a helpful indicator; however, this might not be the case for every type of signal. The autocorrelations of the column vectors of $\boldsymbol{U}$ and $\boldsymbol{V}$ (Fig. 6d) indicate that the first three components contain considerably more information than the remaining ones. Together, the information from the singular values and the autocorrelations suggests that the histogram time series can be adequately described with the first three components, *i.e.*, three species, whereas the remaining components are likely to contain only experimental noise.

What information about the identified species is contained in the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$? As illustrated by Henry & Hofrichter [16], the histograms of all species can contribute to all columns of $\boldsymbol{U}$, and the concentration time courses of all of the species can contribute to every column of $\boldsymbol{V}$. Thus, $\boldsymbol{U}$ and $\boldsymbol{V}$ in principle do contain the information we are looking for, *i.e.*, $\boldsymbol{F}$ and $\boldsymbol{C}$, but unfortunately, this information is not available directly, and the decomposition is not unique. A self-modelling approach has been employed to derive the signal characteristics of the pure species and their concentration time courses [47]; however, the result requires essentially the same choices as in MCR-ALS when it comes to deciding which signature to assume for the pure species. Nevertheless, also without the self-modelling approach, the columns of $\boldsymbol{U}$ contain information about the shape of the species histograms. In our example, the first two columns of $\boldsymbol{U}$ contain features of all three molecular species, while the third only shows features of one species, but with a higher contribution of noise (Fig. 6e). This example illustrates that the columns of $\boldsymbol{U}$ are usually a combination of all species present in the sample and do not result in pure histograms of individual species. However, SVD indicates at which transfer efficiencies changes take place in the histograms and informs us about the timescale of these changes, as illustrated in Fig. 6f for the first four components of the SVD. The kinetics are contained in the columns of $\boldsymbol{V}$, and since they are a combination of the signal from the different molecular species, a global fit of all significant amplitude vectors is necessary to describe the overall kinetics of the process. In a two-state system, this would directly yield the sum of the two rate coefficients of the reaction, but in a more complex system, conclusions regarding the underlying rate coefficients are more difficult to draw. The apparent rates are, however, a convenient way of identifying the relevant timescales in a process under different reaction conditions, without requiring model assumptions.

After deciding which components of the SVD are considered significant and which contain mostly noise, we can reconstruct the data based on the significant components by setting the singular values of all other components to zero (resulting in $\boldsymbol{S}'$) and then calculating $\boldsymbol{H}' = \boldsymbol{U}\boldsymbol{S}'\boldsymbol{V}^T$. Ideally, $\boldsymbol{H}'$ contains all relevant information of the original data while excluding the measurement noise (Fig. 6b), another useful application of SVD. These noise-filtered data can then be used for subsequent fitting procedures [15].

Moreover, reconstructing the data with ***S***' containing different numbers of non-zero values provides a way of assessing the identified number of significant components by comparison with the original data.

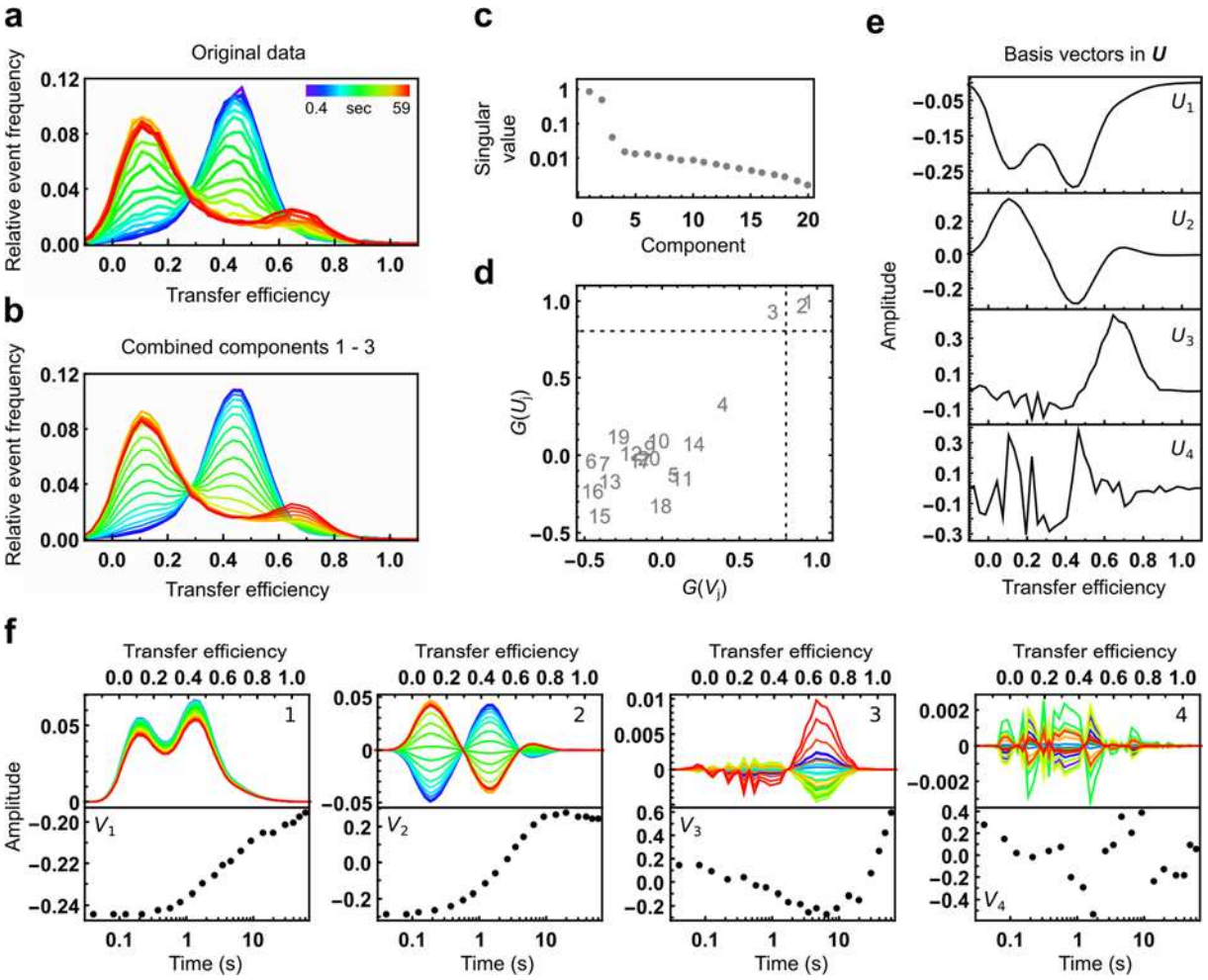

Figure 6. Example of SVD applied to a time series of transfer efficiency histograms. (**a**) Experimentally observed transfer efficiency histograms, normalized to an area of one. Each line represents one transfer efficiency histogram, with the associated time color-coded from purple to green to red (see colour scale). The data show the first minute of ClyA protomer formation measured in a microfluidic mixing device [2]. (**b**) Reconstruction of the data from the first three SVD components yields a representation with reduced noise. (**c**) Singular values (diagonal elements of ***S***) of the SVD of the data shown in a. The first two components are clearly separated from the remaining singular values, indicating their dominance of the signal. (**d**) Autocorrelations of the columns of ***V*** vs. the autocorrelations of the columns of ***U*** of the SVD of the data shown in **a**. The correlation value is indicated by the number of the component as a marker. The result indicates that the third component still contains useful information since it is above the 0.8 threshold (dashed lines, see text), at least for $G(U_j)$. (**e**) The first four columns of ***U***. U_1 to U_3 show clear features expected from signal, but U_4 is dominated by noise. (**f**) Time evolution of the first four SVD components. The top panels show the basis vectors multiplied by their respective singular values and amplitude vectors. The corresponding amplitude vectors (columns of ***V***, see text) are displayed in the lower panels. The time course of the fourth basis vector (V_4) exhibits much more scatter than the previous three, as expected for a basis vector dominated by noise.

In the case of species that contribute only little to the overall signal, SVD can fail to identify their contribution to the data. This is illustrated in Fig. 7, where the transfer efficiency histogram time series [5] clearly shows the presence of three species (Fig. 7a). However, the singular values resulting from SVD show only two significant components (Fig. 7b), as do the autocorrelations of the columns of ***U*** and ***V*** (Fig. 7c). Inspection of the columns of ***U*** shows that the signal contribution of the third species is distributed over several basis vectors and has a smaller amplitude than the noise on species 1 and 2 (Fig. 7d). The latter is the likely cause of the contribution of the third species not being identified by the SVD. As a result, SVD is not useful for the analysis of this minor population, even though its presence and kinetics are highly reproducible [5].

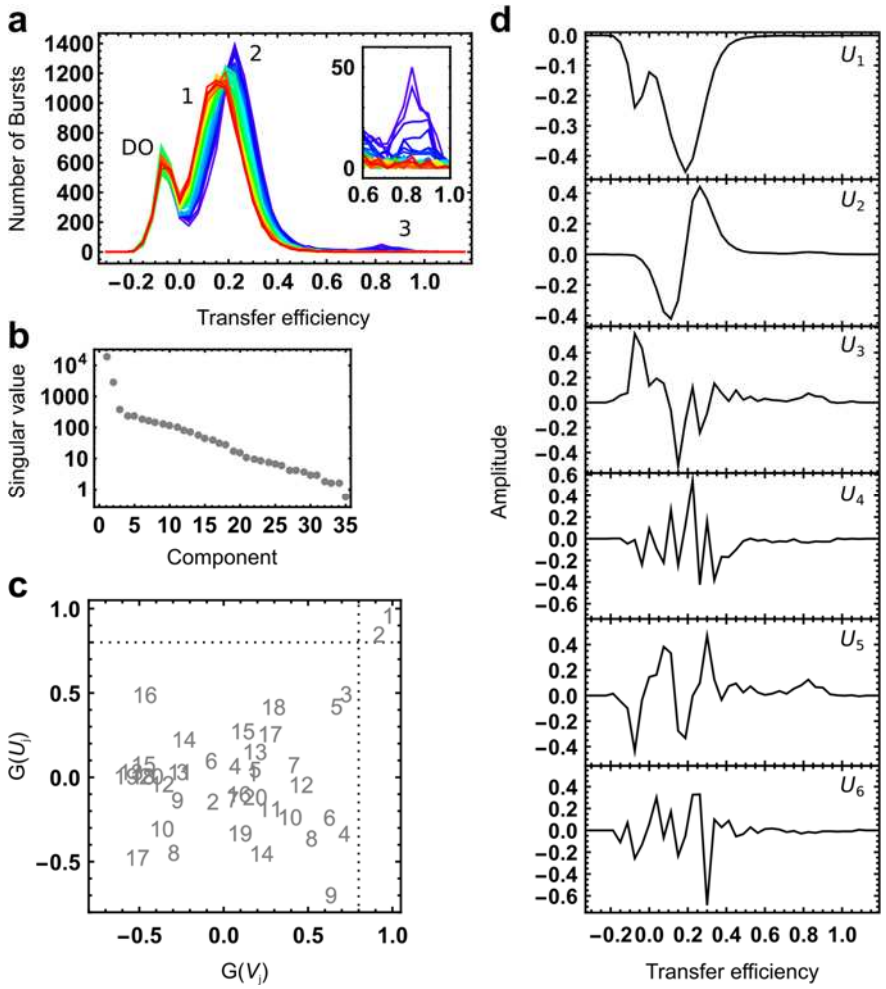

Figure 7. SVD may not detect minor species even if they are significant. (**a**) Transfer efficiency histogram time series from an I27 refolding experiment [5] with numbers indicating the peaks of the three different states. Time progresses from blue to red, with a total time of 1 h for the entire series. The peak labelled 'DO' results from 'donor-only' molecules with inactive or absent acceptor fluorophore, or from fluorescent impurities in the sample. The inset shows a magnification of the small population 3. (**b**) Singular values of the SVD of the data in (a). (**c**) Autocorrelations of the columns of $\boldsymbol{V}$ vs. the autocorrelations of the columns of $\boldsymbol{U}$ of the SVD of the data shown in **a**. The dashed lines indicate the 0.8 threshold. (**d**) First six column vectors of $\boldsymbol{U}$.

5.2. Multiparameter SVD

Biomolecular reactions such as conformational changes in proteins are intrinsically multidimensional. Correspondingly, rates measured with one experimental observable, such as transfer efficiency, do not necessarily coincide with those obtained with other observables. Single-molecule fluorescence experiments offer the opportunity to record multiple observables simultaneously [42]. Even though the transfer efficiency between a donor and an acceptor fluorophore is well suited for monitoring conformational changes in proteins, other quantities such as fluorescence lifetimes or fluorescence anisotropies may contain additional information, in particular if the local environment of the fluorophores differs between the molecular species. For example, fluorescence anisotropies report on the rotational freedom of the fluorophores, whereas fluorescence lifetimes contain information about distance distributions but are also sensitive to processes such as dynamic quenching, which depend on the local environment. Multiparameter single-molecule spectroscopy using polarization sensitive time-correlated single photon counting (TCSPC) with four detection channels can allow the simultaneous acquisition of at least nine different quantities [42]. These quantities include transfer efficiency (E),

acceptor and donor fluorescence anisotropies (r_A, r_D), polarization and fluorophore-dependent fluorescence lifetimes $(\tau_{A\parallel}, \tau_{D\parallel}, \tau_{A\perp}, \tau_{D\perp})$, burst duration (t_b), and photon detection rates. In general, these parameters define a multi-dimensional histogram. However, for most practical purposes, two-dimensional maps with one coordinate being the transfer efficiency and the second coordinate being one of the eight remaining observables are easier to visualize and interpret. A considerable advantage of SVD is the ease with which to include all observables in a global analysis. Instead of one transfer efficiency histogram per time, t_n, a series of two-dimensional histograms will be used. To include these maps in one SVD, each two-dimensional map can be expressed as a vector

$$\boldsymbol{h}_{E,X,n} = \left(h_{1,1,n} \cdots h_{1,K,n},\ h_{2,1,n} \cdots h_{2,K,n}, \cdots, h_{M,1,n} \cdots h_{M,K,n}\right)^{\mathrm{T}}.$$

Here, $h_{m,k,n}$ with $m = 1, \ldots M$ and $k = 1, \ldots K$ is the number of burst events with transfer efficiency values E satisfying $E_m - \Delta/2 < E \leq E_m + \Delta/2$ and observable X satisfying $X_k - \Delta/2 < X \leq X_k + \Delta/2$. In a second step, the vectors $\boldsymbol{h}_{E,X,n}$ for each two-dimensional map are combined in one vector of vectors

$$\boldsymbol{h}_n = \begin{pmatrix} \boldsymbol{h}_{E,r_D,n} \\ \boldsymbol{h}_{E,r_A,n} \\ \vdots \\ \boldsymbol{h}_{E,w,n} \end{pmatrix}.$$

As explained above, the different vectors for all times are combined into the data matrix $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots \boldsymbol{h}_N)$, which can be decomposed into $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{S}$. As in one-dimensional SVD, the vectors of $\boldsymbol{V}$ report on the kinetics, and the components of $\boldsymbol{U}$ contain the two-dimensional basis vectors.

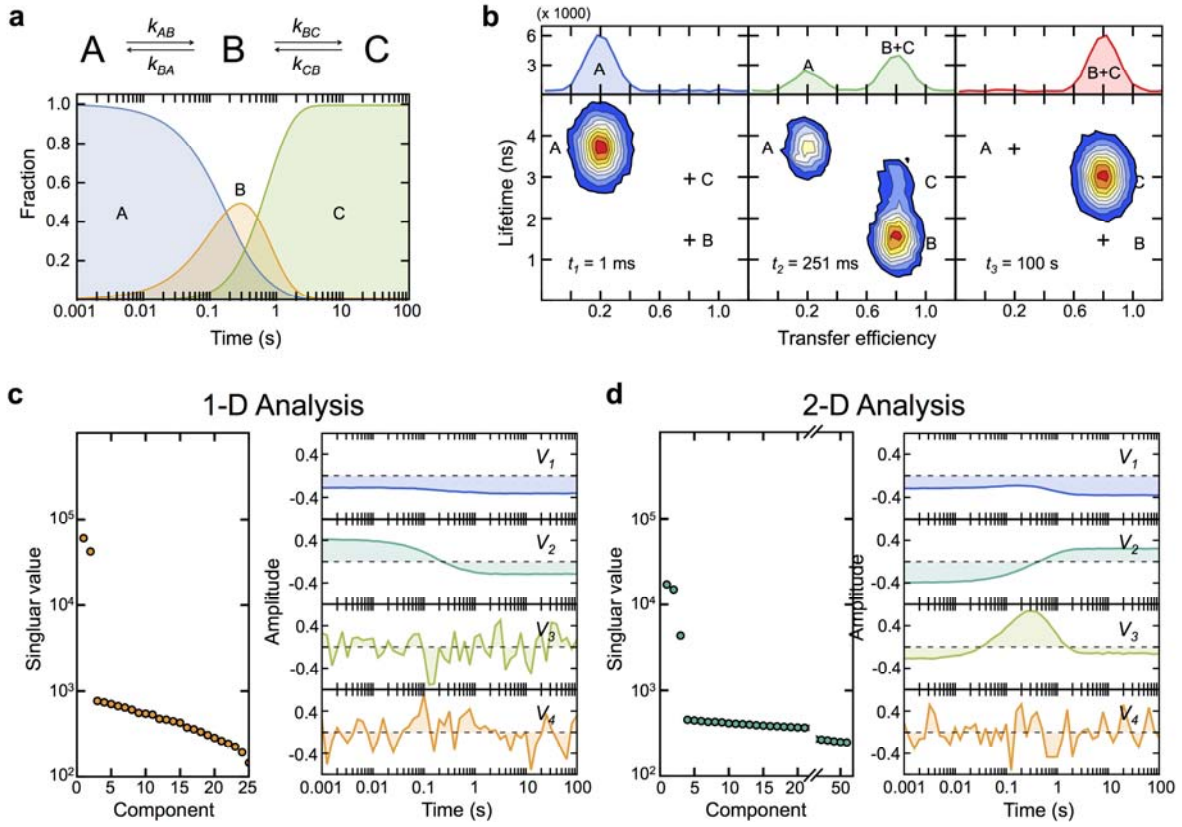

Figure 8. Comparison between one- and two-dimensional SVD on a synthetic dataset. (**a**) Three-state model (top) and resulting kinetics for $k_{AB} = 5\ s^{-1}$, $k_{BA} = 1\ s^{-1}$, $k_{BC} = 2\ s^{-1}$, $k_{CB} = 0.01\ s^{-1}$. (**b**) Two-dimensional histograms using donor fluorescence lifetime versus transfer efficiency, and projections onto the transfer efficiency coordinate (top). (**c**) One-

dimensional SVD with singular values (left) and amplitude vectors (right) using only the projections onto the transfer efficiency. (**d**) Two-dimensional SVD with singular values (left) and amplitude vectors (right).

To illustrate the potential of multiparameter SVD, we compare a one- and two-dimensional synthetic data set of a sequential reaction that involves three molecular species (Fig. 8a). In this particular case, we assume that two of the molecular species (B and C) have identical transfer efficiencies but different fluorescence lifetimes of the donor, *e.g.*, due to dynamic quenching of the donor in state B. Two-dimensional histograms at different times of the reaction from A to C clearly reflect all three species, whereas the projection along the transfer efficiency coordinate only detects two species, one species with low FRET (A) and the other with high FRET (B+C) (Fig. 8b). A comparison of the singular values resulting from one- and two-dimensional SVD shows that we only obtain three significant components if both coordinates, *i.e.*, transfer efficiency and lifetime, are included. If only the transfer efficiency is used as an observable, only two components are significant (Fig. 8c,d). Even though closer inspection of the amplitude vectors, *i.e.*, the components of ***V***, reveals that the decays are not single-exponential, thus indicating the presence of more than two species, the use of an additional variable clearly increases the sensitivity of SVD and thus the identification of kinetic species.

Correspondingly, multiparameter SVD has previously been used to investigate complex reactions in heterogeneous environments, such as the folding of a protein (*bovine* rhodanese) in the cavity formed by the chaperonin GroEL-GroES [20]. In this example, the time series of transfer efficiency histograms shows broad distributions that result from the restricted mobility of the fluorophores in the small cavity formed by the chaperonin (Fig. 9a). These broad and nearly featureless histograms complicate a model-dependent analysis based on peak fitting and call for a less model-dependent analysis such as SVD. Surprisingly, even multidimensional SVD of a data set comprising nine observables resulted in just two significant components with exponential relaxation (Fig. 9b,c), with the same kinetics as those of the transfer efficiency histograms alone. In this case, multidimensional SVD thus provides additional support for a description of the process in terms of only two molecular species.

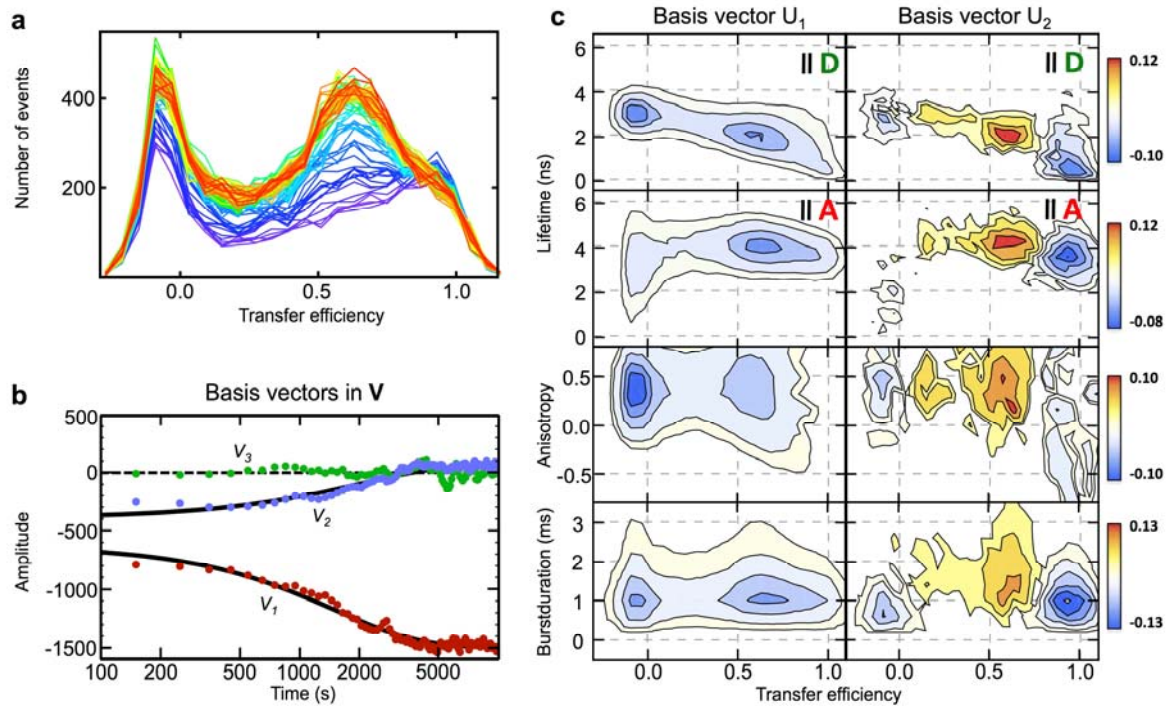

Figure 9. Multidimensional SVD of protein folding in the chaperonin GroEL-GroES. (**a**) Transfer efficiency histograms of rhodanese folding inside the GroEL-GroES cavity as a function of time (*progressing from blue to red*). (**b**) First (red), second (blue), and third (green) amplitude vector weighted by the corresponding singular values of the multi-dimensional SVD for the reaction shown in a. The first two amplitude vectors dominate the observed signal change. Solid lines are a global fit with

a single-exponential function. (**c**) Examples of two-dimensional histograms constructed from the columns of $\boldsymbol{U}$ from multidimensional SVD for the folding reaction in (a) (from Top to Bottom: donor and acceptor fluorescence lifetime, donor fluorescence anisotropy, duration of bursts). The colour code indicates the absolute SVD amplitude (see colour scale). The basis vectors U_2 indicate the positions of changes of the corresponding observables in the data. Data taken from [20].

In summary, SVD enables a model-free analysis of the data and provides information on the minimum number of species involved in a reaction and the timescales of its overall kinetics, which is often useful for further analysis steps. However, SVD does not yield the species histograms or the time courses of their relative concentration that would be necessary for a detailed kinetic analysis, *i.e.*, it is usually not possible to fit and compare kinetic models to the data directly. For a thorough kinetic analysis, this information needs to be extracted from the SVD result with methods resembling those of MCR-ALS and related approaches [15, 47]. We note that the use of multiple parameters, as demonstrated here for SVD, can in principle also be incorporated in the other analysis methods presented here.

6. General experimental considerations

For a stringent kinetic analysis using the methods described here, the experimental data have to meet the following three requirements.

(1) All changes in the relative histogram amplitudes must result from changes in concentrations of the species. Amplitude changes resulting from differences in time binning or acquisition times must be eliminated by normalizing the histograms to the total number of bursts,

$$\sum_{m=1}^{M} h_{ml} = 1.$$

(2) The burst identification algorithm employed must detect the bursts of all species with the same probability, *i.e.*, each species must have the same likelihood to contribute to the transfer efficiency histograms, so that the relative concentrations in $\boldsymbol{C}$ describe the actual relative concentrations of the species present in the reaction. Strictly speaking, this requirement is only met for species with equal diffusion properties and molecular brightness, or if adequate corrections can be applied for the differences. In practice, small differences in brightness and diffusion coefficients can often be neglected for the analysis, but it is important to test for a possible bias during burst identification. For example, changes in the thresholds for burst identification should not influence the relative amplitudes of the peaks in the transfer efficiency histograms. Otherwise, the relative number of bursts identified for the species needs to be corrected. In the case of brightness differences, the absolute brightness of each species has to be determined to correct for the detection bias. Current methods for absolute brightness determination, *e.g.*, FIDA [26] and PDA [25], can be applied if measurements of the isolated species are feasible.

(3) All data should be recorded with identical instrument settings. Combining measurements acquired under different conditions, *e.g.*, changes in instrument alignment, into one data set can require individual sets of species histograms, $\boldsymbol{F}$, for each set of experimental conditions.

7. Conclusion

Single-molecule FRET experiments allow biomolecular reactions to be investigated in great detail. Here we have presented methods for the analysis of time series of transfer efficiency histograms that take full advantage of this opportunity. SVD enables a model-free decomposition of the data and an identification of the minimum number of species in the reaction, the parts of the histograms where changes occur, and

the overall kinetics of the process. However, SVD *per se* does not yield the individual species histograms and the time evolution of their relative concentrations, which are necessary for discriminating between different kinetic models. MCR-ALS requires the number of species as an input but in return provides estimates of the species histograms and concentration time courses. While we can fit kinetic models to the MCR-ALS result, the interdependence of kinetics and the shapes of species histograms can result in ambiguities that can complicate the quantitative comparison of kinetic models. To test and compare different kinetic models based on the histogram time series, model-based peak fitting, including a complete model of peak shapes as well as kinetics usually provides the most reliable strategy. However, SVD and MCR-ALS can be very useful for the process of model building. Finally, we note that the same approaches used here for the analysis of kinetic measurements can of course be employed for equilibrium single-molecule transfer efficiency histograms, *e.g.*, upon systematic variation in solution conditions, such as pH, salt concentration, temperature or the like.

Acknowledgements

This work was supported by the Swiss National Science Foundation.

References

[1] Antonik M, Felekyan S, Gaiduk A and Seidel C A M 2006 Separating structural heterogeneities from stochastic variations in fluorescence resonance energy transfer distributions via photon distribution analysis *J. Phys. Chem. B* **110** 6970-8

[2] Benke S, Roderer D, Wunderlich B, Nettels D, Glockshuber R and Schuler B 2015 The assembly dynamics of the cytolytic pore toxin ClyA *Nat. Commun.* **6** 6198

[3] Borgia A, Kemplen K R, Borgia M B, Soranno A, Shammas S, Wunderlich B, Nettels D, Best R B, Clarke J and Schuler B 2015 Transient misfolding dominates multidomain protein folding *Nat. Commun.* **6** 8861

[4] Borgia A, Wensley B G, Soranno A, Nettels D, Borgia M, Hoffmann A, Pfeil S H, Lipman E A, Clarke J and Schuler B 2012 Localizing Internal Friction along the Reaction Coordinate of Protein Folding by Combining Ensemble and Single Molecule Fluorescence Spectroscopy *Nat. Commun.* **2** 1195

[5] Borgia M B, Borgia A, Best R B, Steward A, Nettels D, Wunderlich B, Schuler B and Clarke J 2011 Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins *Nature* **474** 662-5

[6] Dahan M, Deniz A A, Ha T J, Chemla D S, Schultz P G and Weiss S 1999 Ratiometric measurement and identification of single diffusing molecules *Chem. Phys.* **247** 85-106

[7] de Juan A, Maeder M, Martinez M and Tauler R 2000 Combining hard- and soft-modelling to solve kinetic problems *Chemometrics and Intelligent Laboratory Systems* **54** 123-41

[8] Förster T 1948 Zwischenmolekulare Energiewanderung und Fluoreszenz *Annalen der Physik* **6** 55-75

[9] Frank G A, Goomanovsky M, Davidi A, Ziv G, Horovitz A and Haran G 2010 Out-of-equilibrium conformational cycling of GroEL under saturating ATP concentrations *Proc Natl Acad Sci U S A* **107** 6270-4

[10] Gambin Y, Vandelinder V, Ferreon A C, Lemke E A, Groisman A and Deniz A A 2011 Visualizing a one-way protein encounter complex by ultrafast single-molecule mixing *Nat Methods* **8** 239-41

[11] Gopich I V 2012 Likelihood functions for the analysis of single-molecule binned photon sequences *Chem. Phys.* **396** 53-60

[12] Gopich I V and Szabo A 2005 Theory of photon statistics in single-molecule Förster resonance energy transfer *J. Chem. Phys.* **122** 1-18

[13] Gopich I V and Szabo A 2007 Single-molecule FRET with diffusion and conformational dynamics *J. Phys. Chem. B* **111** 12925-32

[14] Hanwell M D, Curtis D E, Lonie D C, Vandermeersch T, Zurek E and Hutchison G R 2012 Avogadro: an advanced semantic chemical editor, visualization, and analysis platform *J Cheminform* **4** 17

[15] Henry E R 1997 The use of matrix methods in the modeling of spectroscopic data sets *Biophys. J.* **72** 652-73

[16] Henry E R and Hofrichter J 1992 Singular Value Decomposition - Application to Analysis of Experimental-Data *Methods Enzymol.* **210** 129-92

[17] Hoffmann A, Kane A, Nettels D, Hertzog D E, Baumgärtel P, Lengefeld J, Reichardt G, Horsley D A, Seckler R, Bakajin O and Schuler B 2007 Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy *Proc. Natl. Acad. Sci. USA* **104** 105-10

[18] Hoffmann A, Nettels D, Clark J, Borgia A, Radford S E, Clarke J and Schuler B 2011 Quantifying heterogeneity and conformational dynamics from single molecule FRET of diffusing molecules: recurrence analysis of single particles (RASP) *Phys Chem Chem Phys* **13** 1857-71

[19] Hofmann H, Hillger F, Delley C, Hoffmann A, Pfeil S H, Nettels D, Lipman E A and Schuler B 2014 Role of denatured-state properties in chaperonin action probed by single-molecule spectroscopy *Biophys. J.* **107** 2891-902

[20] Hofmann H, Hillger F, Pfeil S H, Hoffmann A, Streich D, Haenni D, Nettels D, Lipman E A and Schuler B 2010 Single-molecule spectroscopy of protein folding in a chaperonin cage *Proc. Natl. Acad. Sci. USA* **107** 11793-8

[21] Jaumot J, de Juan A and Tauler R 2015 MCR-ALS GUI 2.0: New features and applications *Chemometrics and Intelligent Laboratory Systems* **140** 1-12

[22] Jaumot J, Gargallo R, de Juan A and Tauler R 2005 A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB *Chemometrics and Intelligent Laboratory Systems* **76** 101-10

[23] Jaumot J, Marchan V, Gargallo R, Grandas A and Tauler R 2004 Multivariate curve resolution applied to the analysis and resolution of two-dimensional [1H,15N] NMR reaction spectra *Anal Chem* **76** 7094-101

[24] Kalinin S, Felekyan S, Valeri A and Seidel C A 2008 Characterizing multiple molecular states in single-molecule multiparameter fluorescence detection by probability distribution analysis *J Phys Chem B* **112** 8361-74

[25] Kalinin S, Valeri A, Antonik M, Felekyan S and Seidel C A 2010 Detection of structural dynamics by FRET: a photon distribution and fluorescence lifetime analysis of systems with multiple states *J. Phys. Chem. B* **114** 7983-95

[26] Kask P, Palo K, Ullmann D and Gall K 1999 Fluorescence-intensity distribution analysis and its application in biomolecular detection technology *Proc Natl Acad Sci U S A* **96** 13756-61

[27] Knight J B, Vishwanath A, Brody J P and Austin R H 1998 Hydrodynamic focusing on a silicon chip: mixing nanoliters in microseconds *Phys. Rev. Lett.* **80** 3863-6

[28] Kuzmenkina E V, Heyes C D and Nienhaus G U 2005 Single-molecule Forster resonance energy transfer study of protein dynamics under denaturing conditions *Proc. Natl. Acad. Sci. USA* **102** 15471-6

[29] Lee D D and Seung H S 1999 Learning the parts of objects by non-negative matrix factorization *Nature* **401** 788-91

[30] Lipman E A, Schuler B, Bakajin O and Eaton W A 2003 Single-molecule measurement of protein folding kinetics *Science* **301** 1233-5

[31] McKinney S A, Joo C and Ha T 2006 Analysis of single-molecule FRET trajectories using hidden Markov modeling *Biophys. J.* **91** 1941-51

[32] Mueller M, Grauschopf U, Maier T, Glockshuber R and Ban N 2009 The structure of a cytolytic alpha-helical toxin pore reveals its assembly mechanism *Nature* **459** 726-30

[33] Nir E, Michalet X, Hamadani K M, Laurence T A, Neuhauser D, Kovchegov Y and Weiss S 2006 Shot-noise limited single-molecule FRET histograms: Comparison between theory and experiments *J. Phys. Chem. B* **110** 22103-24

[34] Pettersen E F, Goddard T D, Huang C C, Couch G S, Greenblatt D M, Meng E C and Ferrin T E 2004 UCSF chimera - A visualization system for exploratory research and analysis *Journal of Computational Chemistry* **25** 1605-12

[35] Pfeil S H, Wickersham C E, Hoffmann A and Lipman E A 2009 A microfluidic mixing system for single-molecule measurements *Rev. Sci. Instrum.* **80** 055105

[36] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 2007 *Numerical Recipes*: Cambridge University Press)

[37] Rigler R and Elson E S 2001 *Flourescence Correlation Spectroscopy: Theory and Applications* (Berlin: Springer)

[38] Sakmann B and Neher E 1995 *Single Channel Recording*: Plenum Press)

[39] Schuler B and Hofmann H 2013 Single-molecule spectroscopy of protein folding dynamics-expanding scope and timescales *Curr. Opin. Struct. Biol.* **23** 36-47

[40] Schuler B, Lipman E A and Eaton W A 2002 Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy *Nature* **419** 743-7

[41] Selvin P R and Ha T 2008 *Single-Molecule Techniques: A Laboratory Manual* (New York: Cold Spring Harbor Laboratory Press)

[42] Sisamakis E, Valeri A, Kalinin S, Rothwell P J and Seidel C A M 2010 Accurate Single-Molecule FRET Studies Using Multiparameter Fluorescence Detection *Methods Enzymol.* **475** 455-514

[43] Soranno A, Buchli B, Nettels D, Müller-Späth S, Cheng R R, Pfeil S H, Hoffmann A, Lipman E A, Makarov D E and Schuler B 2012 Quantifying internal friction in unfolded and intrinsically disordered proteins with single molecule spectroscopy *Proc. Natl. Acad. Sci. USA* **109** 17800-6

[44] Wallace A J, Stillman T J, Atkins A, Jamieson S J, Bullough P A, Green J and Artymiuk P J 2000 E-coli hemolysin E (HlyE, ClyA, SheA): X-ray crystal structure of the toxin and observation of membrane pores by electron microscopy *Cell* **100** 265-76

[45] Wunderlich B, Nettels D, Benke S, Clark J, Weidner S, Hofmann H, Pfeil S H and Schuler B 2013 Microfluidic mixer designed for performing single-molecule kinetics with confocal detection on timescales from milliseconds to minutes *Nat. protocols* **8** 1459-74

[46] Xie Z, Srividya N, Sosnick T R, Pan T and Scherer N F 2004 Single-molecule studies highlight conformational heterogeneity in the early folding steps of a large ribozyme *Proc Natl Acad Sci U S A* **101** 534-9

[47] Zimanyi L 2004 Analysis of the bacteriorhodopsin photocycle by singular value decomposition with self-modeling: A critical evaluation using realistic simulated data *Journal of Physical Chemistry B* **108** 4199-209